



UNIVERSIDADE FEDERAL DE SANTA CATARINA
Centro de Ciências da Educação
CURSO DE GRADUAÇÃO EM BIBLIOTECONOMIA



RAFAEL ROECK BORGES CASSETTARI

**LEI DE ZIPF EM DISCURSOS ORAIS: UMA COMPARAÇÃO ENTRE
TRABALHOS ACADÊMICOS ESCRITOS E SUAS APRESENTAÇÕES**

Florianópolis
2014

Rafael Roeck Borges Cassettari

Lei de Zipf em discursos orais: uma comparação entre trabalhos acadêmicos escritos e suas apresentações

Trabalho de Conclusão do Curso de Graduação em Biblioteconomia, do Centro de Ciências da Educação da Universidade Federal de Santa Catarina, requisito parcial à obtenção do título de Bacharel em Biblioteconomia.

Orientador: Dr. Adilson Luiz Pinto

Florianópolis
2014

Ficha catalográfica elaborada pelo graduando de Biblioteconomia/UFSC Rafael
Roeck Borges Cassettari

C344l Cassettari, Rafael, 1991-
Lei de Zipf em discursos orais: uma comparação entre
trabalhos acadêmicos escritos e suas apresentações /
Rafael Cassettari. – Florianópolis, 2014.
40 f. ; 30 cm

Orientador: Dr. Adilson Luiz Pinto.
Trabalho de conclusão de curso (graduação em
Biblioteconomia) – Centro de Ciências da Educação,
Universidade Federal de Santa Catarina, 2014.

1. Lei de Zipf. 2. Bibliometria 3. Estatísticas Linguísticas
I. Título

CDD 025.04

Esta obra é licenciada por uma licença Creative Commons de atribuição, de uso não comercial
e de compartilhamento pela mesma licença 2.5



 creative
commons

Você pode:

- copiar, distribuir, exibir e executar a obra;
- criar obras derivadas.

Sob as seguintes condições:

- Atribuição. Você deve dar crédito ao autor original.
- Uso não-comercial. Você não pode utilizar esta obra com finalidades comerciais.
- Compartilhamento pela mesma licença. Se você alterar, transformar ou criar outra obra com base nesta, somente poderá distribuir a obra resultante com uma licença idêntica a esta.

Acadêmico: Rafael Roeck Borges Cassettari

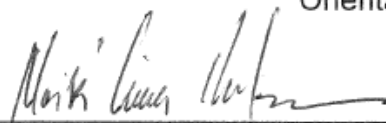
Título: Lei de Zipf em discursos orais: uma comparação entre trabalhos acadêmicos escritos e suas apresentações.

Trabalho de conclusão do curso de Biblioteconomia, do centro de Ciências da Educação da Universidade Federal de Santa Catarina, como requisito parcial para obtenção do grau de Bacharel em Biblioteconomia, aprovada com a nota 100

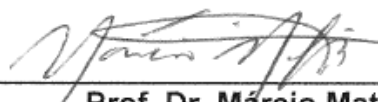
Florianópolis, 09 de julho de 2014.



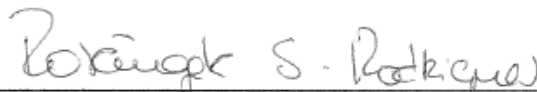
Prof. Dr. Adilson Luiz Pinto
Universidade Federal de Santa Catarina
Orientador



Prof. Dr. Moisés Lima Dutra
Universidade Federal de Santa Catarina
Membro da Banca Examinadora



Prof. Dr. Márcio Matias
Universidade Federal de Santa Catarina
Membro da Banca Examinadora



Profa. Dra. Rosângela Schwarz Rodrigues
Universidade Federal de Santa Catarina
Membro Suplente

RESUMO

CASSETTARI, Rafael. **Lei de Zipf em discursos orais**: uma comparação entre trabalhos acadêmicos escritos e suas apresentações. 2014. 38f. TCC (Graduação) – Curso de Biblioteconomia, Departamento de Ciências da Educação, Universidade Federal de Santa Catarina, 2014.

A lei de Zipf é uma teoria com base na matemática e linguística que analisa e quantifica como as palavras são distribuídas dentro de um determinado texto. Desta forma, é possível representar por meio de gráficos e análises estatísticas quais são os termos que mais se repetem de modo que seja possível criar um ranking de palavras-chave. Esta pesquisa verificou, por meio da lei de Zipf, as variações e uniformidades de trabalhos acadêmicos escritos (artigos) e os mesmos apresentados de forma oral. As apresentações orais foram inseridas em forma de vídeo no Youtube, para que fosse possível recuperar, de forma automática, a transcrição do áudio. Por meio de um script executado em Bash, os textos e as apresentações transcritas foram quantificadas e organizadas, sendo possível criar nuvens de tags e tabelas com os rankings, facilitando a análise entre os conteúdos. Foi possível identificar as esferas dos conteúdos, identificar as palavras em comum ou muito distantes e analisar e comparar matematicamente o que foi escrito com o que foi apresentado.

Palavras-chave: Lei de Zipf. Bibliometria. Estatísticas linguísticas.

ABSTRACT

CASSETTARI, Rafael. **Lei de Zipf em discursos orais**: uma comparação entre trabalhos acadêmicos escritos e suas apresentações. 2014. 38f. TCC (Graduação) – Curso de Biblioteconomia, Departamento de Ciências da Educação, Universidade Federal de Santa Catarina, 2014.

Zipf's law is a theory based on mathematics and linguistics that analyzes and quantifies how words are distributed within a text. It is possible to represent by graphs and statistical analyzes which are the terms that are repeated over so that a ranking of keywords is created. This research found, through the Zipf's law, variations and uniformities of written academic papers and they presented orally. The oral presentations were inserted in video form on Youtube, it was possible to recover automatically the transcript of the audio. Using a Bash script, texts and transcribed presentations were quantified and organized, thereby creating tag clouds and tables with rankings, facilitating the analysis of the contents. It was possible to identify the spheres of content, identifying common words or not and, mathematically, analyze and compare what was written with what was presented.

Keywords: Zipf's Law. Bibliometrics. Linguistics statistics.

LISTA DE FIGURAS

Figura 1 – Nuvem de palavras do Texto 1.....	19
Figura 2 – Nuvem de palavras do Texto 2.....	20
Figura 3 – Nuvem de palavras do Texto 3.....	21
Figura 4 – Nuvem de palavras da Apresentação 1.....	22
Figura 5 – Nuvem de palavras da Apresentação 2.....	23
Figura 6 – Nuvem de palavras da Apresentação 3.....	24

LISTA DE TABELAS

Tabela 1 – Tabela dos dados extraídos do conteúdo 1 (Textual x Oral).....	26
Tabela 2 – Ranking de palavras do conteúdo 1.....	27
Tabela 3 – Ranking de palavras da apresentação 1.....	28
Tabela 4 – Tabela dos dados extraídos do conteúdo 2 (Textual x Oral).....	29
Tabela 5 – Ranking de palavras do conteúdo 2.....	30
Tabela 6 – Ranking de palavras da apresentação 2.....	31
Tabela 7 – Tabela dos dados extraídos do conteúdo 3 (Textual x Oral).....	32
Tabela 8 – Ranking de palavras do conteúdo 3.....	33
Tabela 9 – Ranking de palavras da apresentação 3.....	34

SUMÁRIO

1 INTRODUÇÃO	9
2 OBJETIVOS	10
2.1 Objetivo geral.....	10
2.2 Objetivos específicos.....	10
3 REFERENCIAL TEÓRICO.....	11
3.1 Bibliometria	11
3.2 Teoria de Zipf	12
3.3 Aplicações da Teoria de Zipf.....	13
4 METODOLOGIA.....	16
5 RESULTADOS	19
5.1 Análise dos conteúdos textuais.....	19
5.2 Análise dos conteúdos orais.....	23
5.3 Análise comparativo entre os conteúdos Textual X Oral	27
6 DISCUSSÃO E CONCLUSÕES.....	37
REFERÊNCIAS	39

1 INTRODUÇÃO

A área de Biblioteconomia, hoje em dia, vai muito além da visão clássica ou popular de tratar apenas de assuntos dentro de uma unidade de informação, mais especificamente a biblioteca. Hoje, abraçando a Ciência da Informação, uma série de estudos que envolvem descobertas relevantes, transformando dado em informação e informação em conhecimento, vários estudos cresceram neste meio.

Na parte de Estatística e Matemática da área da Ciências da Informação, surgiu a bibliometria, que tem como sua principal função analisar a qualidade e quantidade baseados em alguns indicadores e leis do estudo. No ambiente científico, a Bibliometria obteve um destaque para a obtenção de indicadores métricos, de modo que seja possível medir a qualidade e quantidade de materiais científicos. Dentro dos estudos bibliométricos, surgiram várias leis e teorias, sendo as principais a lei de Bradford, a lei de Lotka e a lei de Zipf. Esta última é a lei que move este estudo.

A lei de Zipf é uma base matemática-linguística que analisa a frequência e distribuição das palavras contidas em um texto, seja ele científico ou não. Por meio de um cálculo, é possível mapear e criar *rankings* de ocorrência das palavras neste texto. Após isso, cabe ao analista interpretar os dados coletados.

Neste estudo, foi aplicada a lei de Zipf em textos escritos e textos orais, mais especificamente em trabalhos acadêmicos apresentados em congresso. Em um primeiro momento, foram analisados os textos escritos e, paralelamente, os textos orais. Por fim, feito o cruzamento e interpretação destes dados a fim de procurar semelhanças ou diferenças nos tipos apresentados.

Previamente, foram levantadas algumas questões para a criação desta pesquisa que envolve a comparação entre o que é escrito e o que é falado. O que é dito pelo indivíduo condiz com o que foi escrito? Utilizamos maneiras diferentes de expressar uma mesma coisa? A apresentação oral tem a mesma importância que o trabalho acadêmico escrito? Partindo dessas perguntas, este estudo irá analisar, aplicando a lei de Zipf, três textos escritos e suas respectivas apresentações orais.

2 OBJETIVOS

2.1 Objetivo geral

Analisar os trabalhos escritos e orais apresentados em um congresso a fim de identificar as semelhanças e diferenças entre os conteúdos.

2.2 Objetivos específicos

- a) Identificar quais são as três esferas da lei de Zipf de cada texto, de modo que seja possível criar um *ranking* de frequência das palavras encontradas;
- b) Identificar as palavras comuns entre as apresentações orais e escritas;
- c) Comparar a posição de cada termo, para que seja possível analisar as principais diferenças entre o que foi escrito e o que foi dito;
- d) Comparar e analisar, matematicamente, as diferenças entre os textos.

3 REFERENCIAL TEÓRICO

A bibliometria cresceu e se destacou no meio científico por conta na atual necessidade de se ter valores e atributos que definam a qualidade e a quantidade de material existente.

Essa necessidade foi primeiramente percebida no século XX, quando algumas pesquisas sobre estudos métricos começaram a aparecer. Em 1926, a primeira lei mais conhecida destes estudos surgiu, a lei de Lotka, tendo como principal foco do seu estudo analisar a produção dos pesquisadores. No ano de 1934, Bradford, com o seu método de análise da concentração das publicações. Em 1949, a lei de Zipf que trabalha principalmente com a análise da frequência e distribuição das palavras em um determinado texto.

3.1 Bibliometria

Segundo Araújo (2006) o termo “bibliometria” sofreu uma série de mudanças e discussões em questão de nome e de entendimento da sua função. O termo, de início, era conhecido por “bibliografia estatística” em 1923 por Wyndham Hulme e, somente anos depois, em 1934, Paul Otlet, usou o termo “bibliometria”. Após uma série de discussões, no final da década de 60 “bibliometria” acabou se popularizando.

A partir da análise e entendimento das três leis clássicas da bibliometria, é possível ter uma maior compreensão da sua função e aplicação.

A primeira lei clássica é a lei de Lotka, onde o Araújo (2006) relata que o estudo de Lotka foi sobre a produtividade dos pesquisadores que publicavam na revista “Chemical Abstracts” nos anos de 1909 e 1916 e, após feita esta análise, percebeu que a grande parte da publicação científica está concentrada em um pequeno grupo de cientistas e que, inversamente, um grande grupo de pesquisadores detêm apenas uma pequena parte da produção.

Bradford foi quem criou o segundo estudo que deu origem à segunda lei clássica da bibliometria. Este estudo era focado em um conjunto de periódicos científicos, diferentemente de Lotka que estudou, especificamente, os autores. Com este estudo, Bradford relatou que, os periódicos que têm uma maior concentração de uma publicação de determinado assunto são os periódicos importantes daquele

tema. De modo matemático, todos os artigos analisados devem ser somados e dividido por três, sendo que o primeiro terço desta divisão é a parte principal deste assunto. Com este valor, analisa-se quantas revistas entram nesta fração. Os periódicos dentro desta fatia é considerado importante para aquele assunto.

Zipf é o criador da lei clássica da bibliometria e que, de acordo com Rousseau e Zhang (1992), foi o estudo que tentou explicar a ocorrência de palavras em um texto baseado no princípio do menor esforço. “De acordo com este princípio, pessoas tentam achar um equilíbrio entre a uniformidade e a diversidade no uso das palavras” (ROUSSEAU e ZHANG, 1992, p.2, tradução nossa). Desta maneira, ainda como os autores Rousseau e Zhang trazem, é como quando não lembramos o nome de alguém e acabamos substituindo o nome da pessoa por outros termos e que no âmbito científico ocorre do mesmo jeito. Cientistas acabam trocando algumas terminologias de modo que busque um equilíbrio no seu vocabulário. Esta será a base de nosso estudo.

3.2 Teoria de Zipf

O estudo de Zipf trabalha com a frequência de aparição de uma palavra em um texto, independentemente do que esta simbologia represente sozinha ou em conjunto. Por exemplo a palavra “a”, em um contexto geral, significa muito mais que apenas a primeira letra de nosso alfabeto mas, nesta lei, ela é apenas mais uma palavra dentro de um longo texto.

Os autores Chen e Leimkuhler (1987) nos mostram a importância da lei de Zipf afirmando que ela serve para estudar os mais diversos tipos de fenômenos humanos, montando *rankings* que exibem, de uma maneira matemática, esses fenômenos que podem ser analisados de diversas formas.

Por meio desta base matemática, é possível traçar uma série de fenômenos apenas analisando a frequência em que determinadas palavras aparecem em um texto.

A base matemática da sua lei é:

$$\sqrt{TP} = PE \quad \sqrt{PE} = SE$$

“TP” é o Total de Palavras (ocorrências); “PE”, Primeira Esfera; “SE”, Segunda Esfera”.

Ou seja, a raiz quadrada do total de palavras nos dá a primeira esfera. A primeira esfera é, então, as palavras que atingem, pelo menos, a raiz quadrada do total de palavras. Por exemplo, se o total de palavras é 625, a primeira esfera são as palavras que têm, pelo menos, 25 (vinte e cinco) ocorrências. A segunda esfera, é a raiz quadrada da primeira esfera, no caso, 5 (cinco). A terceira esfera, que não aparece na fórmula, seria o restante das palavras, aquelas as que aparecem menos que cinco vezes.

Deste modo, com a lei de Zipf, é possível traçar um *ranking* de palavras que mais aparecerem em um determinado trecho ou texto.

Como destaca Günter e outros autores (1996), os dados e o *ranking* criados nada mais são que variáveis aleatórias. É necessário entender o motivo desta ordem de palavras e analisar sua significância perante o estudo.

Ainda no pensamento de Günter, entender o porquê destas palavras aparecerem na ordem descoberta é o principal motivo da utilização da teoria de Zipf. Esta ordem está diretamente ligada com a importância dos termos que aparecem mais vezes dentro do texto.

Há várias formas de se analisar o texto utilizando a lei de Zipf e também várias formas de produto final com ela como, por exemplo, a análise de palavras-chave.

3.3 Aplicações da Teoria de Zipf

A teoria de Zipf pode ser aplicada de diversas maneiras em análises de linguística. A sua principal e mais básica aplicação são em textos escritos, onde, como explicado anteriormente, é feita uma análise sobre as ocorrências das palavras.

É possível aliar a lei de Zipf em estudos de recuperação da informação, como Quoniam e outros autores (1998) destacam, analisando as ocorrências em comparação às palavras-chave de determinado texto. Esta teoria é utilizada para qualificar todas as palavras presentes no texto e, assim, com o seu ranking, entender o assunto do texto ou até comparar se as palavras-chaves selecionadas pelo autor têm realmente alguma ligação.

Segundo um estudo de Paliwal, Bhatnagar e Haldar (1986), a lei de Zipf foi utilizada para traçar a previsão da utilização de determinados recursos, baseando-se na repetição em que cada recurso aparece. Desta forma, é possível mapear e listar estes recursos, focar que material deve ser mais explorado (no caso de minérios extraídos, por exemplo), entre outras interpretações.

Uma das aplicações da lei em questão é para analisar o índice/quantidade de co-citações, citações diretas ou indiretas e até determinadas vezes em que um autor é citado dentro de um artigo utilizando as fórmulas onde, baseado na repetição do nome do autor, seja possível encontrar estes índices. Um exemplo disso é o estudo de Ruiz-Baños e outros autores (1999), que utilizam a teoria bibliométrica de Zipf para quantificar e criar um modelo de presença de co-citações.

Outra forma possível de utilização da lei de Zipf é, que será utilizado em nosso estudo, a análise em discursos orais. A diferença é que não há necessariamente um texto escrito. No caso, há uma transcrição do discurso oral para ser feita a análise. A base matemática é a mesma, porém a análise final deve ser um pouco mais cuidadosa, uma vez que o que é dito muitas vezes carregam 'vícios' ou expressões não-formais ou até mesmo regionalistas, que, se não observado corretamente, pode comprometer a análise.

Ridley (1982) escreveu sobre a lei de Zipf focada para os discursos orais. Ele explica que várias palavras irão aparecer muitas vezes, e elas são, por exemplo, "o", "os", "a" e "as", pois são artigos que, em nosso idioma, são frequentemente utilizados em um texto, independentemente de ser extenso ou não e, por outro lado, palavras que são pouco utilizadas, como é o caso de palavras compostas sobre termos específicos como 'autodestruição' ou 'superestimado'. Porém, não devem ser descartadas as diferenças psicológicas em um texto oral e um texto escrito, mesmo que o assunto deles sejam o mesmo, os indivíduos tendem a utilizar um vocabulário diferente ao escrever ou falar.

O estudioso Garner (1997) também escreveu sobre a lei de Zipf em textos orais. Ele ressalta que muitas palavras ditas tendem a se repetir mais, mesmo que tenham outros sinônimos. O exemplo dado é a palavra 'ready', que em inglês significa 'pronto' ou 'certo', esta palavra tem um significado de "entendimento" e que poderia ser facilmente trocada por palavras como 'okay' ou 'right'. Em um texto

escrito, optamos sempre por diferenciar as palavras escritas, para dar uma dimensão de vocabulário mais extenso ou até mesmo em uma questão de estética. Já em apresentações orais, esta questão estética é deixada de lado em termos comuns, como o 'ready'.

No Português, isso também ocorre, porém a estrutura e o jeito de escrever pode nos mostrar valores bem diferentes do que de outra língua, como o Inglês. Fazer uma análise em nosso idioma, é, em partes, uma tarefa mais difícil, por conta das conjugações verbais. A frase 'Eu amo você' e 'Nós amamos você', aparece o verbo 'Amar', porém, de duas maneiras diferentes. Já no inglês, seria 'I love you' e 'We love you', onde o verbo que no infinitivo é 'to love' aparece duas vezes de maneira igual. Neste caso, a diferença é que, em uma contagem final sobre uma análise fictícia sobre estes textos, no inglês, a palavra apareceria duas vezes e no português, seriam duas diferentes, porém com mesmo sentido. Outro exemplo são os artigos definidos que no português são quatro e no inglês apenas um.

4 METODOLOGIA

Este estudo foi uma análise de estudo de caso, ou seja, uma análise do discurso oral da apresentação de um trabalho com o texto do próprio trabalho.

Em nosso pré-teste, foram escolhidos três vídeos, que estão no Youtube, divididos em três categorias e um tema pré-definidos. Estas categorias foram separadas pelo tempo de execução do vídeo. 01 vídeo com duração menor que 5 minutos; 01 vídeo com duração entre 5~15 minutos; 01 vídeo com duração maior que 15 minutos. O tema escolhido foi a tag do youtube “Bibliometrics” (bibliometria em português). Dado o tema e a divisão pelo tempo de execução do vídeo, foram escolhidos 03 vídeos de forma aleatória, apenas respeitando as restrições da pesquisa.

No estudo final, por conta do resultado deste pré-teste, foi definido um novo foco do estudo onde seria feita a análise de um discurso oral da apresentação de um trabalho de congresso, que foi filmado e colocado no Youtube, em modo privado, apenas para a coleta das legendas automáticas, com o próprio texto do trabalho publicado.

O Youtube tem uma ferramenta chamada 'Automatic Closed Captions' ou 'Legendas automáticas'. Com esta função, é possível que um vídeo em inglês seja transcrito. Nem todas as palavras são 'interpretadas' de maneira correta, há uma quantidade razoável de erros desta ferramenta na língua portuguesa, porém, em inglês, a precisão das legendas automáticas é altíssima. A coleta foi feita de maneira automática mas, após a geração dos arquivos, foram feitas conferências para correção de alguns erros.

O Software 'Youtube-DL', é um software, gratuito, escrito na linguagem de programação 'Python'. Python é uma linguagem de alto nível orientada a objetos. É uma linguagem dinâmica, modular e multiplataforma. Foi criada pelo holandês Guido van Rossum, em 1991. O Python é usado de modo geral, pois suas bibliotecas incluem diversas ferramentas desde simples processamentos de texto até utilização de serviços do sistema operacional.

Este software é capaz de baixar integralmente vários dados de um determinado vídeo no Youtube. Uma das funções é a possibilidade de baixar as

legendas que estão inseridas no vídeo. Com essa função, será possível pegar toda a transcrição dos vídeos selecionados e fazer a análise do que foi falado neste vídeo.

Será utilizado um script executado em Bash para separar e quantificar os dados coletados pela ferramenta Youtube-DL.

Scripts são um conjunto de instruções de uma determinada linguagem de programação de modo que seja feita uma extensão de sua funcionalidade. Estas instruções são montadas em forma de algoritmo com uma sequência lógica das funções onde, de acordo com as respostas do usuário, o *script* segue um determinado fluxo do algoritmo.

Bash é um interpretador de comandos em que é possível executar uma série de comandos diretos do sistema operacional. Estes comandos mexem diretamente com operações aritméticas (linguagem das máquinas).

Com os dados coletados (falas em formato texto), é possível executar o script para quantificar os dados. Nas próprias instruções, os dados gerados serão repassados para um arquivo de extensão '.CSV'. 'Comma-Separated Values' (Valores separados por vírgula), é um formato caracterizado como 'ordenador de bytes'. Como o nome sugere, as informações são separadas por vírgula e, utilizando o interpretador certo, os dados serão tabulados automaticamente a partir dos padrões separados pela vírgula.

O comando em bash será o seguinte: `cat LEGENDAS | tr -cs '[:lower:][:upper:]' '[n*]' | sort | uniq -c | sed 's/^ *\[0-9*\] \^1,/ ' > ARQUIVO.csv`. Este comando, em partes, fará as seguintes etapas:

“cat LEGENDAS”: “cat” é o comando de leitura de arquivos texto.;

“tr -cs '[:lower:][:upper:]' '[n*]’”: “tr” é o comando que copia as entradas principais dos caracteres e repassa para uma nova saída. O parâmetro “-cs '[:lower:][:upper:]' '[n*]’” irá transformar todos os caracteres de entrada de modo padronizado, inclusive as repetições;

“sort” - irá organizar as palavras coletadas;

“uniq -c” - remove as palavras repetidas e adiciona, ao lado dela, a quantidade de vezes que esta palavra foi repetida no arquivo;

“sed 's/^ *\[0-9]*\)/1,/' > ARQUIVO.csv” - irá criar um arquivo .CSV, com os dados já refinados, em ordem alfa-numérica do modo “Palavra, número de repetições”.

Como a língua portuguesa, diferentemente da inglesa, possui caracteres especiais (acentos, por exemplo), utilizamos outro comando para a remoção dos acentos. Algumas palavras perdem o sentido por conta da sua acentuação, porém em uma grande quantidade de palavras, poucas se somariam tanto à chegar em um ponto que se tornasse um problema, como por exemplo a palavra ‘publica’ e ‘pública’.

O comando utilizado previamente nos arquivos de texto é:

“sed 'y/áÀàÃãÂâÉéÊêÍíÓóÔôÕõÚúÇ/aAaAaAaEeEiloOoOoOuUcC/' ARQUIVO_DE_ENTRADA.txt > ARQUIVO_DE_SAÍDA.txt”

Este comando irá substituir todos os acentos (primeira parte do comando) pelas letras 'limpas' (segunda parte do comando).

Em nosso estudo, iremos tratar as esferas da lei de Zipf do seguinte modo:

- Primeira Esfera: Palavras mais comuns dentro do texto. São incluídos artigos definidos e indefinidos, preposições, conectores, etc.
- Segunda Esfera: Palavras mais importantes dentro do texto. São as palavras que ocorreram em menor número que as palavras citadas da primeira esfera. Por não serem palavras de uso comum, serão consideradas como importantes.
- Terceira Esfera: Palavras consideradas como ruídos dentro do texto, sem qualquer importância.

Com o auxílio de um software que gerencie planilhas, como o 'BrOffice Calc' (disponível gratuitamente), os dados coletados foram analisados e comparados para obter o resultado final.

5 RESULTADOS

As análises foram feitas de maneira individual e coletiva. Serão apresentadas as análises em todos os textos orais e nos textos escritos com o seu texto oral relativo, de maneira que seja possível interpretar as correlações entre um texto oral com outro apresentado e também encontrar as semelhanças e diferenças entre a apresentação e o texto publicado.

Na primeira parte, será feita uma análise individual de cada texto.

Na segunda parte, a análise será entre cada texto escrito e sua apresentação oral.

Os textos orais tiveram, em média, 25 minutos de duração.

5.1 Análise dos conteúdos textuais

O primeiro texto tem como seu título “Os livros nas teses da ciência da informação: um estudo de citação”.

Este trabalho teve um total de 1226 palavras diferentes em um total de 4379 palavras escritas. A palavra com o maior número de ocorrências foi 'a' com 202 repetições. Aplicando a fórmula de Zipf, encontramos as três esferas divididas nos seguintes valores:

- Primeira esfera: 48 palavras (até 14 repetições);
- Segunda esfera: 213 palavras (até 3 repetições);
- Terceira esfera: 966 palavras (restante).

Figura 1 - Nuvem de palavras do Texto 1



Fonte: elaborada pelos autores.

Na primeira esfera cabe destacar as palavras 'citação' e 'citações' que aparecem 81 e 23 vezes respectivamente e que a palavra “ciência” e “informação” aparecem 56 e 41 vezes respectivamente. É possível perceber a importância que estas palavras tem dentro do seu escopo, uma vez que estes termos se encontram, inclusive, no título do trabalho.

Em sua segunda esfera, as palavras que tem o maior destaque são 'artigos', 'científico', 'fontes' e 'informação'. Também em uma análise partindo do título do texto, a importância destas palavras também devem ser destacadas, uma vez que elas são os termos que dão início à segunda esfera, que é considerada interessante.

O texto seguinte leva o título “Redes de coautoria em ciência da informação: estudo sobre a natureza das relações na área temática de 'organização e representação do conhecimento”.

O trabalho escrito contém 1190 palavras em um total de 4753 palavras escritas no total. O maior número de ocorrências ficou com a palavra 'de' que repetiu 311 vezes. A divisão das esferas ficou:

- Primeira esfera: 52 palavras (até 17 repetições);
- Segunda esfera: 123 palavras (até 4 repetições);
- Terceira esfera: 1009 palavras (restante).

passando para a nona posição. Outra palavra que deve ser destacada é a palavra 'coletânea' que junta 17 aparições.

Na segunda parcela, 107 protagonizam o que chamaríamos de 'parte interessante' onde as palavras que dominam este topo são 'científico', 'perfil', 'saber' e 'sentido', todas com 10 ocorrências. A palavra 'tese' aparece em 8 oportunidades e que, se somada com a palavra 'teses' que apareceu na primeira esfera, somariam 29 e ocupariam a décima posição. Outras palavras interessantes que possuem grande relação com o texto são 'pragmática', 'relação', 'documento', 'autores', 'artigo', com 6 ocorrências e 'científica' e 'semântica', com 5 repetições.

O texto seguinte apresentam um total de 660 palavras e 2809 ocorrências e a palavra com mais ocorrências foi 'de' em 188 oportunidades. As esferas foram divididas no seguinte modo:

- Primeira esfera: 42 palavras (até 13 repetições);
- Segunda esfera: 134 palavras (até 3 repetições);
- Terceira esfera: 484 palavras (restante).

Figura 5 - Nuvem de palavras da Apresentação 2



Fonte: elaborada pelos autores.

Na primeira divisão há uma palavra que merece uma atenção especial por se tratar de um prefixo. A palavra 'co' aparece 101 vezes sendo que, analisando o texto separadamente, percebe-se que ela aparece sempre em duas oportunidades. Ou aparece como 'co-citação(ões)' ou como 'co-autores/autoria'. A palavra 'citação'

aparece 68 vezes e a palavra 'citações' aparece 14 vezes, dando um total de 82 repetições. A palavra 'autores' aparece 56 vezes e a palavra 'autoria' 14 vezes, onde a soma é 70. Analisando em uma maneira geral, sem qualquer análise individual no texto, a soma dessas ocorrências é de 152 aparições e que 101 vezes elas aparecem juntas do prefixo 'co', que tem grande ligação com o título e contexto do trabalho apresentado.

Ainda na primeira esfera, há outras palavras que devem ser destacadas como 'citados', que aparece 26 vezes, 'frequência', 20 vezes, 'Salton', 13 vezes. A palavra 'Salton' aparece na primeira esfera por se tratar no 'Cosseno de Salton' onde, inclusive, a palavra 'cosseno' aparece na segunda esfera com 10 repetições.

O terceiro e último texto traz um total de 577 palavras únicas e 2326 repetições. A palavra que apareceu mais foi 'e' 128 vezes. A divisão das esferas foram:

- Primeira esfera: 44 palavras (até 11 repetições);
- Segunda esfera: 133 palavras (até 3 repetições);
- Terceira esfera: 400 palavras (restante);

5.3 Análise comparativo entre os conteúdos Textual X Oral

O primeiro conteúdo ficou dividido nas seguintes informações:

Tabela 1: **Tabela de dados extraídos do conteúdo 1 (Textual X Oral)**

Identificação de conteúdos	Textual	Oral
<i>Palavras únicas</i>	1226	571
<i>Palavras totais</i>	4379	2064
<i>Palavras na primeira esfera</i>	48	32
<i>Repetições dentro da primeira esfera</i>	14	11
<i>Palavras na segunda esfera</i>	213	107
<i>Repetições dentro da segunda esfera</i>	13 entre 3	10 entre 3

Fonte: elaborada pelos autores.

Esta primeira tabela serve para termos uma ideia dos valores coletados de uma maneira individual, para, após, levantar algumas conclusões.

É possível, por meio desta tabela, ver a relação entre as palavras escritas e as palavras ditas. Alguns resultados:

- A relação Escrito/Oral das palavras únicas é de 2,14:1 e de palavras totais é de 2,12:1
- A média entre as palavras totais e as palavras únicas no texto escrito é de 3,57 palavras;
- A média entre as palavras totais e as palavras únicas no texto oral é de 3,61 palavras;
- A primeira esfera do trabalho escrito representa 3,91% das palavras únicas;
- A primeira esfera do trabalho oral representa 5,6% das palavras únicas
- A segunda esfera do trabalho escrito representa 17,37% das palavras únicas;
- A segunda esfera do trabalho oral representa 18,73% das palavras únicas;

Neste primeiro texto já é possível perceber a semelhança proporcional entre a apresentação oral e o trabalho escrito.

A tabela abaixo mostra as 10 primeiras palavras da primeira esfera do texto escrito junto com o seu *Ranking* e aparição nos textos orais e escritos. Nesta tabela foram excluídos os artigos, preposições e conectores, porém foi mantida a posição do *ranking*. O 'critério de desempate' utilizado é a organização alfanumérica.

Tabela 2: **Ranking de palavras do conteúdo 1**

Palavra	Número de aparições no texto escrito	Posição no Ranking escrito	Número de aparições no texto oral	Posição no Ranking oral
<i>Citações</i>	81	6 ^a	16	25 ^a
<i>Livros</i>	73	8 ^a	17	23 ^a
<i>Coletâneas</i>	44	17 ^a	17	21 ^a
<i>Ciência</i>	41	19 ^a	17	20 ^a
<i>Informação</i>	41	20 ^a	32	9 ^a
<i>Zona</i>	37	22 ^a	X	X
<i>Teses</i>	36	24 ^a	21	15 ^a
<i>Citados</i>	32	27 ^a	7	48 ^a (SEGUNDA ESFERA)
<i>Autores</i>	31	28 ^a	6	53 ^a (SEGUNDA ESFERA)
<i>Citação</i>	23	34 ^a	23	12 ^a

Fonte: elaborada pelos autores.

Algumas informações relevantes retiradas da tabela 2:

- As palavras com maior ocorrência no texto escrito tem uma posição inversa no texto oral;
- A palavra 'informação' é a que tem a melhor posição no texto oral, ocupando a 9^a posição mas, no texto escrito, é apenas a quinta mais importante, ocupando a 20^a posição;
- A palavra 'citação' obteve a maior posição no texto escrito, ocupando a 6^a posição e no texto oral ocupa apenas a 25^a posição;
- Entre as 10 primeiras palavras da primeira esfera do texto escrito, apenas 7 também se encontram na primeira esfera do texto oral;
- A palavra 'Zona' é a sexta palavra com mais aparições no texto escrito, ocupando a 22^a opção, porém ela não foi dita uma única vez no texto oral. A palavra 'zonas' foi dita apenas uma vez.

A próxima tabela é uma análise feita da maneira contrária à anterior. São as 10 primeiras palavras da primeira esfera do texto oral.

Tabela 3: **Ranking de Palavras da Apresentação 1**

<i>Termo</i>	<i>Número de aparições no texto escrito</i>	<i>Posição no Ranking escrito</i>	<i>Número de aparições no texto oral</i>	<i>Posição no Ranking oral</i>
<i>Informação</i>	41	20 ^a	32	9 ^a
<i>Livro</i>	16	41 ^a	24	11 ^a
<i>Citação</i>	24	34 ^a	23	12 ^a
<i>Teses</i>	36	24 ^a	21	15 ^a
<i>Ciência</i>	41	19 ^a	17	20 ^a
<i>Coletâneas</i>	44	17 ^a	17	21 ^a
<i>Livros</i>	73	8 ^a	17	23 ^a
<i>Citações</i>	81	6 ^a	16	25 ^a
<i>Campo</i>	29	30 ^a	14	29 ^a
<i>Científico</i>	12	51 ^a (SEGUNDA ESFERA)	10	30 ^a

Fonte: elaborada pelos autores.

A tabela 3 mostra a situação inversa onde:

- As palavras 'livro', 'campo' e 'científico' não aparecem no ranking das 10 palavras que mais se repetem no texto escrito;
- A palavra 'campo' tem uma importância semelhante tanto no texto escrito como no texto oral;
- Apenas a palavra 'científico' não está na primeira esfera do texto escrito.

A tabela abaixo mostra as informações gerais do segundo texto analisado.

Tabela 4: Tabela de dados extraídos do conteúdo 2 (Textual X Oral)

	<i>Escrito</i>	<i>Oral</i>
<i>Palavras únicas</i>	1190	660
<i>Palavras totais</i>	4753	2809
<i>Palavras na primeira esfera</i>	52	42
<i>Repetições dentro da primeira esfera</i>	17	13
<i>Palavras na segunda esfera</i>	123	134
<i>Repetições dentro da segunda esfera</i>	16~4	12~3

Fonte: elaborada pelos autores.

Com base na tabela acima é possível destacar que:

- A relação Escrito/Oral das palavras únicas é de 1,8:1 e de palavras totais é de 1,69:1
- A média entre as palavras totais e as palavras únicas no texto escrito é de 3,99 palavras;
- A média entre as palavras totais e as palavras únicas no texto oral é de 4,25 palavras;
- A primeira esfera do trabalho escrito representa 4,36% das palavras únicas;
- A primeira esfera do trabalho oral representa 6,3% das palavras únicas
- A segunda esfera do trabalho escrito representa 10,33% das palavras únicas;
- A segunda esfera do trabalho oral representa 20,3% das palavras únicas;

Como no texto anterior, este trabalho tem uma semelhança grande entre a relação de palavras únicas e totais em ambos formatos de apresentação. A proximidade entre as proporções da primeira esfera também é visível.

A próxima tabela mostra as 10 primeiras palavras da primeira esfera do texto escrito junto com o seu *Ranking* e aparição nos textos orais e escritos do segundo trabalho analisado.

Tabela 5: Ranking de palavras do Texto 2

<i>Termo</i>	<i>Número de aparições no texto escrito</i>	<i>Posição no Ranking escrito</i>	<i>Número de aparições no texto oral</i>	<i>Posição no Ranking oral</i>
<i>Relações</i>	84	5 ^a	2	270 ^a (TERCEIRA ESFERA)
<i>Relação</i>	53	11 ^a	5	102 ^a (SEGUNDA ESFERA)
<i>Redes</i>	48	14 ^a	3	163 ^a (SEGUNDA ESFERA)
<i>Pesquisa</i>	45	16 ^a	3	151 ^a (SEGUNDA ESFERA)
<i>Orientador Científica</i>	39	19 ^a	X	X
	36	20 ^a	6	79 ^a (SEGUNDA ESFERA)
<i>Nível</i>	36	21 ^a	X	X
<i>Orientando</i>	36	22 ^a	X	X
<i>Mestrado</i>	35	24 ^a	X	X
<i>Doutorado</i>	30	28 ^a	X	X

Fonte: elaborada pelos autores.

As informações que podemos retirar da análise da tabela 5 é que:

- Do mesmo modo que o trabalho analisado anteriormente, as posições são inversas no texto escrito e oral;
- Não houve nenhuma ocorrência em que as 10 primeiras palavras da primeira esfera no texto escrito estivessem no texto oral;
- 5 termos que mais se repetiram no trabalho escrito não tiveram qualquer aparição na apresentação oral;

A tabela 6 faz o ranking inverso do anterior, pontuando os 10 termos que tiveram mais repetições no texto oral.

Tabela 6: Ranking de palavras na Apresentação 2

Termo	Número de aparições no texto escrito	Posição no Ranking escrito	Número de aparições no texto oral	Posição no Ranking oral
Citação	1	593 ^a (TERCEIRA ESFERA)	68	6 ^a
Autores	3	191 ^a (SEGUNDA ESFERA)	56	8 ^a
Análise	8	87 ^a (SEGUNDA ESFERA)	38	11 ^a
Citados	X	X	26	17 ^a
Frequência	3	218 ^a (SEGUNDA ESFERA)	20	22 ^a
Domínio	1	716 ^a (TERCEIRA ESFERA)	18	26 ^a
Proximidade	1	1051 ^a (TERCEIRA ESFERA)	17	30 ^a
Matriz	X	X	15	32 ^a
Autoria	9	75 ^a (SEGUNDA ESFERA)	14	33 ^a
Citações	1	594 ^a (TERCEIRA ESFERA)	14	34 ^a

Fonte: elaborada pelos autores.

Algumas informações de destaque:

- A lógica da disposição do ranking ser inverso continua;
- Não há palavras da primeira esfera oral na primeira esfera escrita;
- Duas palavras não foram mencionadas no texto escrito e outras quatro foram apenas mencionadas uma única vez;
- Do mesmo modo que a análise do texto escrito, o texto oral utiliza outras palavras para ser apresentado;
- Com base apenas nestes dados, é possível afirmar que o texto oral e escrito não estão alinhados com a ideia principal. Como o trabalho escrito

foi criado antes, pode-se dizer que a apresentação oral não condiz com o que foi apresentado na forma escrita;

As próximas tabelas são dados do terceiro texto analisado.

Tabela 7: Tabela de dados extraídos do conteúdo 3 (Textual X Oral)

	<i>Escrito</i>	<i>Oral</i>
<i>Palavras únicas</i>	928	577
<i>Palavras totais</i>	4834	2326
<i>Palavras na primeira esfera</i>	48	44
<i>Repetições dentro da primeira esfera</i>	17	11
<i>Palavras na segunda esfera</i>	166	133
<i>Repetições dentro da segunda esfera</i>	16~4	10~3

Fonte: elaborada pelos autores.

Com as informações da tabela 7 é possível dizer que:

- A relação Escrito/Oral das palavras únicas é de 2,07:1 e de palavras totais é de 1,6:1
- A média entre as palavras totais e as palavras únicas no texto escrito é de 5,2 palavras;
- A média entre as palavras totais e as palavras únicas no texto oral é de 4,03 palavras;
- A primeira esfera do trabalho escrito representa 5,17% das palavras únicas;
- A primeira esfera do trabalho oral representa 7,62% das palavras únicas
- A segunda esfera do trabalho escrito representa 17,88% das palavras únicas;
- A segunda esfera do trabalho oral representa 23,05% das palavras únicas;

Utilizando como base os textos anteriores, o terceiro texto segue alguns dados que podemos afirmar que acabam 'se padronizando' em uma apresentação oral de um trabalho escrito, onde a relação de palavras são sempre valores bem próximos (de modo proporcional).

A tabela 8 traz o ranking das 10 palavras que mais se repetem no terceiro texto escrito junto com os dados relativos da apresentação oral.

Tabela 8: **Ranking de palavras do Texto 3**

<i>Termo</i>	<i>Número de aparições no texto escrito</i>	<i>Posição no Ranking escrito</i>	<i>Número de aparições no texto oral</i>	<i>Posição no Ranking oral</i>
<i>Páginas</i>	82	6 ^a	3	157 ^a (SEGUNDA ESFERA)
<i>Citações</i>	81	7 ^a	14	33 ^a
<i>Universidades</i>	66	11 ^a	31	10 ^a
<i>Conexões</i>	55	16 ^a	6	72 ^a
<i>UFRGS</i>	51	18 ^a	24	16 ^a
<i>UFSC</i>	41	21 ^a	17	25 ^a
<i>Websites</i>	39	22 ^a	17	26 ^a
<i>Lugar</i>	37	23 ^a	2	232 ^a (TERCEIRA ESFERA)
<i>Web</i>	36	25 ^a	7	69 ^a (SEGUNDA ESFERA)
<i>Universidade</i>	34	26 ^a	28	13 ^a

Fonte: elaborada pelos autores.

Informações que devem ser destacadas após a análise da tabela 7:

- A apresentação oral do terceiro texto foi a única que trouxe todas as 10 primeiras palavras da primeira esfera do trabalho escrito onde 2 palavras

se encontram na segunda esfera e 1 na terceira esfera do texto oral.

- O posicionamento (ranking) das palavras da primeira esfera do texto escrito são, relativamente, parecidas na apresentação oral, destacando as palavras 'universidades', 'websites', 'UFRGS' e 'UFSC';
- Diferentemente dos outros textos, não demonstrou uma relação inversa no posicionamento dos termos.

A última tabela apresentada traz a análise inversa, sobre o posicionamento das palavras de maior repetição no texto oral sobre as palavras do texto escrito:

Tabela 9: **Ranking de palavras da Apresentação 3**

<i>Termo</i>	<i>Número de aparições no texto escrito</i>	<i>Posição no Ranking escrito</i>	<i>Número de aparições no texto oral</i>	<i>Posição no Ranking oral</i>
<i>Universidades</i>	66	11 ^a	31	10 ^a
<i>Universidade</i>	34	26 ^a	28	13 ^a
<i>UFRGS</i>	51	18 ^a	24	16 ^a
<i>Página</i>	15	61 ^a (SEGUNDA ESFERA)	19	20 ^a
<i>UFSC</i>	41	21 ^a	17	25 ^a
<i>Websites</i>	39	22 ^a	17	26 ^a
<i>Website</i>	10	97 ^a (SEGUNDA ESFERA)	16	30 ^a
<i>Citações</i>	81	7 ^a	14	33 ^a
<i>Link</i>	6	137 ^a (SEGUNDA ESFERA)	12	39 ^a
<i>Produção</i>	16	54 ^a (SEGUNDA ESFERA)	12	41 ^a

Fonte: elaborada pelos autores.

De acordo com os dados da tabela 9 podemos afirmar que:

- As principais palavras do discurso oral fazem parte do texto escrito onde em 6 oportunidades aparecem na primeira esfera dos formatos apresentados;
- A palavra com menos repetições relativas no texto escrito é 'Website', com 10 repetições. Porém, se levar em consideração que o seu plural 'websites' atingiu a marca de 39 repetições, elevaria o termo a um dos mais utilizados onde apareceria por 49 oportunidades.
- O cenário semelhante detalhado anteriormente também ocorre com a palavra 'Página' em que, no trabalho escrito, não se repete tantas vezes mas no plural é a principal palavra. 'Páginas' aparece 82 no texto escrito.

6 DISCUSSÃO E CONCLUSÕES

Com base nos objetivos traçados no trabalho, observa-se que eles foram atingidos, utilizando a lei de Zipf para a criação de um ranking de palavras que servisse de base para a análise dos textos escritos e orais.

Os resultados detalhados nos trazem a dimensão do trabalho escrito e sua importância dentro do meio acadêmico e que, de um modo geral, a apresentação oral nem sempre traz os detalhes mais importantes que foram descritos na parte escrita. Levando em consideração, principalmente, que no modo escrito não há tantos 'vícios de linguagem' e informalidade que há nas apresentações orais. Também deve-se analisar os fatores humanos que há em uma apresentação, como a interação com os ouvintes, questões fisiológicas (como estar doente, por exemplo), etc.

De uma maneira crua e apenas analisando os números, é possível perceber que o trabalho escrito traz, em proporções, uma base forte para a apresentação oral, onde em média, a cada 2 palavras escritas é dita 1 palavra na apresentação oral de 25 minutos.

Uma possível sugestão de análise futura seria entender o porquê do texto oral fugir, mesmo que de forma leve, do contexto geral do trabalho escrito, uma vez que esta apresentação é baseada, totalmente, no trabalho escrito.

Alguns levantamentos e considerações finais feitos após a leitura e análise de todos os textos:

- O texto escrito não pode ser substituído, de maneira acadêmica, pela apresentação oral;
- As apresentações orais, em sua maioria, buscam palavras diferentes para explicar o que foi feito no trabalho escrito;
- Não há exatamente uma 'preocupação' ao detalhar o trabalho escrito em uma apresentação oral, uma vez que esta análise individual das palavras não é feita. Isso também pode ser levado em consideração e questionada a inserção de palavras-chave no texto;
- A primeira esfera é maior nas apresentações orais, porém, como analisado, nem sempre esta primeira esfera pode ser levada em consideração como 'norteador' do trabalho escrito;

- De modo geral, a segunda esfera no texto oral seria a parte 'mais importante'. Como dito anteriormente, há vícios de linguagem e expressões informais/rotineiras que acabam desvirtuando o sentido principal da primeira esfera e, por isso, estas palavras acabam sendo encontradas apenas na segunda parte.

Esta pesquisa alcançou seus objetivos apresentados onde foram analisados os trabalhos escritos e orais apresentados em um congresso, a fim de identificar semelhanças e diferenças entre suas apresentações, usando a lei de Zipf.

Foram identificadas as três esferas dos textos. Nestas esferas, foi possível criar um ranking e interpretar as palavras relevantes dentro destas esferas. Com a identificação dos grupos, o segundo e terceiro objetivo específico foi facilmente alcançado, sendo possível comparar as posições no ranking e comparar as posições das palavras, como por exemplo uma palavra que apareceu muito em sua apresentação oral mas pouco no texto escrito.

Na primeira esfera, sempre aparecerá em grande número os artigos, preposições e conectores, porém, eles não devem ser descartados uma vez que servirão como base para o cálculo, mesmo que não tenham um “sentido real” dentro do conteúdo.

A análise matemática feita dos textos, trouxe informações relevantes para futuras pesquisas, incluindo a reflexão sobre qual é o real valor acadêmico das apresentações orais em congressos, uma vez que podem haver diferenças grandes entre o que foi escrito e o que foi dito. Deve-se levar em consideração a necessidade da leitura e entendimento do texto e não somente adquirir os dados matematicamente. As estatísticas necessitam de interpretação para conferir os reais significados das palavras inseridas naquela apresentação escrita ou oral.

REFERÊNCIAS

- ARAÚJO, Carlos Alberto. Bibliometria: evolução histórica e questões atuais. **Em Questão**, Porto Alegre, v. 12, n. 1, p.11-32, jan./jun. 2006. Disponível em: <<http://www.brapci.ufpr.br/download.php?dd0=10036>>. Acesso em: mai. 2014.
- CHEN, Ye-Suo; LEIMKUEHLER, Ferdinand F. Analysis of Zipf's law: an index approach. **Information Processing & Management**, Great Britain, v. 23, n. 3, p. 171-182. 1987.
- GARNER, Philip N. On topic identification and dialogue move recognition. **Computer speech and Language**, v. 11, p. 275-306. 1997.
- GÜNTHER, R.; et al. Zipf's law and the effort of ranking on probability distributions. **International Journal of Theoretical Physics**, v. 35, n. 2, p. 395-417. 1996.
- PALIWAL, H. V.; BHATNAGER, S. N.; HALDAR, S. K. Lead-zinc resource prediction in India: an application of Zipf's law. **Mathematical Geology**, v. 18, n. 6, p. 539-549. 1986.
- QUONIAM, L.; et al. Bibliometric law used for information retrieval. **Scientometrics**, Oxford, v. 41, n. 1-2, p. 83-91. 1998.
- RIDDLEY, Dennis R. Zipf's law in transcribed speech. **Psychological Research**, Springer, v. 44, p. 93-103, jan./set. 1982.
- ROUSSEAU, R.; ZHANG, Q. Zipf's data on the frequency of chinese words revisited. **Scientometrics**, Oxford, v. 24, n. 2, p. 201-220. 1992.
- RUIZ-BAÑOS, R.; et al. Structure and dynamics of scientific networks. Part II: the new Zipf's law, the clusters of co-citations and the model of the descriptor presence. **Scientometrics**, Oxford, v. 44, n. 2, p. 235-265. 1999.